Brain Informatics

RESEARCH

Open Access

Explainable brain age prediction: a comparative evaluation of morphometric and deep learning pipelines

Maria Luigia Natalia De Bonis[1†], Giuseppe Fasano[1†], Angela Lombardi[1*], Carmelo Ardito[1], Antonio Ferrara[1], Eugenio Di Sciascio[1] and Tommaso Di Noia[1]

Abstract

Brain age, a biomarker reflecting brain health relative to chronological age, is increasingly used in neuroimaging to detect early signs of neurodegenerative diseases and support personalized treatment plans. Two primary approaches for brain age prediction have emerged: morphometric feature extraction from MRI scans and deep learning (DL) applied to raw MRI data. However, a systematic comparison of these methods regarding performance, interpretability, and clinical utility has been limited. In this study, we present a comparative evaluation of two pipelines: one using morphometric features from FreeSurfer and the other employing 3D convolutional neural networks (CNNs). Using a multisite neuroimaging dataset, we assessed both model performance and the interpretability of predictions through eXplainable Artificial Intelligence (XAI) methods, applying SHAP to the feature-based pipeline and Grad-CAM and DeepSHAP to the CNN-based pipeline. Our results show comparable performance between the two pipelines in Leave-One-Site-Out (LOSO) validation, achieving state-of-the-art performance on the independent test set ($MAE = 3.21$with DNN and morphometric features and $MAE = 3.08$ with a DenseNet-121 architecture). SHAP provided the most consistent and interpretable results, while DeepSHAP exhibited greater variability. Further work is needed to assess the clinical utility of Grad-CAM. This study addresses a critical gap by systematically comparing the interpretability of multiple XAI methods across distinct brain age prediction pipelines. Our findings underscore the importance of integrating XAI into clinical practice, offering insights into how XAI outputs vary and their potential utility for clinicians.

Keywords EXplainable Artificial Intelligence, Brain age prediction, Morphometry, Convolutional neural networks

1 Introduction

Brain age, the estimation of an individual's brain health relative to chronological age, has emerged as a valuable biomarker in neuroimaging studies [1, 2]. Accurate brain age prediction can provide critical insights into the aging process, identify early signs of neurodegenerative diseases, and facilitate the development of personalized treatment plans [3–6]. Indeed, brain age prediction studies have been conducted in a wide range of clinical populations, including neurological conditions such as Alzheimer's disease (AD) [7], mild cognitive impairment (MCI) [8], multiple sclerosis (MS) [9], and traumatic brain injury [1, 10]. These studies aim to assess brain aging in individuals at various stages of health and disease, supporting diagnosis, prognosis, and treatment decisions. Other clinical conditions investigated include epilepsy [11], stroke [12], and psychiatric disorders like schizophrenia [13], bipolar disorder [14], and autism spectrum disorder [15], among others. These studies

†Maria Luigia Natalia De Bonis and Giuseppe Fasano contributed equally to this work.

*Correspondence:
Angela Lombardi
angela.lombardi@poliba.it
[1] Department of Electrical and Information Engineering, Polytechnic University of Bari, Via E. Orabona, 4, 70125 Bari, Italy

underline the broad applicability of brain age prediction as a tool in clinical settings to evaluate neurological and psychiatric conditions, providing insights into disease progression and individualized care. As a result, various machine learning (ML) approaches have been employed to enhance the accuracy and reliability of brain age prediction models [10, 16].

Recent research has focused on two primary methodologies for brain age prediction: the use of morphometric features extracted from MRI scans [17, 18] and the application of deep learning (DL) techniques to raw or minimally preprocessed MRI data [19]. Traditional approaches often involve feature extraction using tools like FreeSurfer[1] to derive morphometric features, which are then fed into machine learning models for age prediction [15]. Conversely, deep learning architectures, particularly three-dimensional convolutional neural networks (3D CNNs), have gained traction for their ability to directly process volumetric T1-weighted MRI scans without requiring extensive preprocessing or intermediate feature extraction steps [1]. Both methodologies have demonstrated promising results, yet they have seldom been compared systematically in terms of performance, interpretability, and clinical usability.

Explainable Artificial Intelligence (XAI) has become increasingly relevant in the context of brain age prediction, as it addresses the black-box nature of many advanced ML and DL models [20, 21]. The goal of XAI is to make model predictions more transparent and understandable, which is particularly important in clinical settings where healthcare professionals must trust and validate decisions [22, 23]. Various XAI methods, such as SHAP (SHapley Additive exPlanations) [24, 25], Grad-CAM (Gradient-weighted Class Activation Mapping) [26], and DeepSHAP [24], offer different mechanisms to interpret model outputs. However, existing literature often employs a single XAI method per class of algorithms and presents aggregated results without delving deeply into the insights provided by different combinations of machine learning methods and XAI techniques.

To bridge the gap between interpretability and trustworthiness in clinical practice, it is crucial to investigate the stability and utility of XAI methods. This involves examining how XAI outputs vary with different parameters and determining if these explanations can be integrated into tools that are practically useful for clinicians. This paper addresses these needs by exploring several research questions:

- RQ1: do different pipelines yield statistically significant differences in performance?
- RQ2: are different XAI methods stable across various parameter settings, and how do they enhance interpretability in the context of these pipelines?
- RQ3: how effectively can the combined insights from pipeline performance and XAI explanations support clinical decision-making in brain age prediction tasks?

To answer these questions, we use a multisite dataset of healthy control groups to compare multiple machine learning architectures within two distinct pipelines-one based on morphometric features and the other one on 3D DNNs trained on minimally preprocessed MRI data. The primary objective of these pipelines is the prediction of brain age, which requires using data exclusively from healthy subjects to establish a normative baseline for chronological age estimation. We select SHAP for the first pipeline and Grad-CAM and DeepSHAP for the second, conducting a thorough analysis of individual explanations across various parameter settings. Additionally, we perform a correlation analysis between the most stable XAI methods and the subjects' ages to identify statistically significant age biomarkers. This work presents a systematic framework for comparative evaluation of the two most adopted brain age prediction pipelines. Unlike isolated statistical comparisons, our framework integrates XAI methods to assess how the explanations provided by different pipelines align with clinical requirements. By offering a unified approach to performance and interpretability assessment, this study contributes to the development of robust and interpretable brain age prediction models for clinical applications.

2 Related works

2.1 Morphometric feature-based pipelines

Morphometric feature-based pipelines leverage anatomical features such as cortical thickness and brain volume extracted from MRI scans to predict brain age. These methods offer a structured approach to understanding brain aging. Cole and colleagues [27] utilized a comprehensive set of neuroimaging phenotypes from the UK Biobank to model brain age, achieving a mean absolute error (MAE) of 3.5 and a correlation coefficient $R = 0.78$. Their study effectively used a broad range of neuroimaging phenotypes to enhance predictive accuracy. Similarly, Madan et al. [28] used different parcellation approaches to predict age from cortical structure, achieving an MAE of 6.5. This comprehensive approach underscored the variability and robustness of cortical measures in brain age prediction. Guan et al. [29] focused on multimodal MRI data, achieving an $MAE = 8.24$ and an $R = 0.85$.

[1] https://surfer.nmr.mgh.harvard.edu/fswiki/recon-all.

Their use of partial least squares regression (PLSR) highlighted specific brain regions significant in predicting brain age. In [30] different feature selection techniques were employed in a multimodal approach, achieving a remarkable MAE of 1.17, demonstrating how focusing on the most relevant features can enhance model performance. In the work of Aycheh et al. [31], the brain age of 2911 healthy subjects was predicted with $MAE = 4.05$ by using cortical thickness data with gaussian process regression (GPR) emphasizing significant cortical features. Lombardi et al. compared different ML models with a set of morphological features, reporting an $MAE = 4.6$ on a hold-out test and underscored the value of detailed feature importance analysis [32].

2.2 DNN pipelines

Deep neural networks excel in brain age prediction by processing raw or minimally preprocessed MRI data. These models capture complex patterns and relationships within the data that traditional methods might miss.

Peng et al. [33] utilized a lightweight CNN architecture with 3D convolutions, achieving an $MAE = 2.58$ and an $R = 0.9$, showcasing the efficiency of CNNs in handling volumetric data. In Dartora et al. [34], different CNN models were compared to verify generalizability with various populations and MRI scan characteristics. In [35], sections of brain-extracted T1-weighted MRI scans were used as input to produce a single scalar regression output by adapting the successful VGG-16 architecture for age regression, achieving an $MAE = 2.9$ and $R = 0.87$. Levakov et al. [36] achieved a mean $MAE = 3.72$, illustrating the challenges of maintaining high accuracy across diverse datasets with 3D CNN architectures.

Besson et al. [37] adapted geometric deep learning techniques, achieving $R = 0.92$ and $MAE = 4.91$. This study demonstrated the versatility of graph CNN in handling complex brain shapes.

2.3 XAI methods and interpretability insights

Interpretability is crucial for clinical applications of brain age prediction models. Various studies have employed different XAI methods to elucidate the decision-making processes of their models.

Among the morphometric feature-based studies, in most of the works, detailed analyses were performed to highlight significant neuroimaging phenotypes and brain regions [27, 29, 30, 32]. These analyses provided clear evidence of specific anatomical features that could serve as biomarkers of aging. In general, among the morphometric feature-based studies, most works focus on feature importance analysis and visualizing the most relevant features for age prediction. Only a smaller portion of studies employ local XAI methods to extrapolate the impact of morphometric features individually [38].

In DNN-based approaches, gradient-based techniques and relevance maps are predominantly used for interpretability. Grad-CAM is a visualization technique that highlights important regions in an input image by using the gradients of the target concept flowing into the final convolutional layer to produce a coarse localization map. This method has been applied to 3D CNNs to visualize which brain regions contribute most to age prediction. Gradient-based methods generate attention maps, providing intuitive visual insights into the specific brain regions the model focused on [39]. Similarly, relevance maps show how much each input feature contributes to the final prediction. Techniques like Layer-wise Relevance Propagation (LRP) can be used to create these maps. In [37] the cortical regions involved in their predictions were identified, using relevance maps to bridge the gap between complex model outputs and clinically meaningful interpretations.

A critical issue with DNN-based XAI methods such as Grad-CAM, and relevance maps is that they often present heatmaps averaged over groups of individuals. While these group-level explanations can highlight common patterns and regions of interest, they do not always translate effectively to individual patient care [36, 40]. In clinical practice, medical specialists need to understand the predictions for individual patients to make informed decisions. Most studies have focused on group-level interpretations without thoroughly exploring the usability of individual-level explanations [36, 37, 41]. The challenge lies in ensuring that these XAI methods can provide reliable and interpretable insights for single patients, which is essential for clinical applications. Individual-level heatmaps must be clear and actionable, helping specialists understand the brain regions influencing each patient's predicted age.

3 Materials

3.1 Dataset

In this study, we used publicly accessible data from OpenBHB.[2] OpenBHB aggregates data from 10 publicly available datasets, including IXI, ABIDE 1, ABIDE 2, CoRR, GSP, LOCALIZER, MPI-Leipzig, NAR, NPC, and RBP. OpenBHB provides data for a specific challenge that includes separate training and validation sets, as well as a private test set [42]. For our study, we exclusively used the publicly accessible data, designating the training set as our training data and the validation set as our

[2] https://ieee-dataport.org/open-access/openbhb-multi-site-brain-mri-dataset-age-prediction-and-debiasing.

Table 1 OpenBHB demographic information

Study	Subjects	Age	Sex (%M)	Sites
ABIDE I	453	17.09 ± 7.85	82.56	20
ABIDE II	462	14.97 ± 9.31	72.29	16
CoRR	600	24.73 ± 15.25	48.33	18
GSP	1342	21.46 ± 2.79	41.73	5
IXI	484	48.52 ± 16.47	45.04	1
Localizer	65	24.06 ± 6.39	46.15	2
MPI-Leipzig	237	34.57 ± 17.32	57.81	1
NAR	251	21.93 ± 4.70	40.24	3
NPC	56	26.32 ± 4.23	46.43	1
RBP	34	23.06 ± 4.99	50.00	1
Total	3984	24.92 ± 14.29	52.38	62

test data. The use of the OpenBHB dataset is particularly justified given the potential biases introduced by imaging characteristics and acquisition sites [15, 38, 43]. Indeed, scanner differences, acquisition protocols and individual characteristics in the training set, such as sex and age distribution, can significantly affect the generalizability of machine learning models trained on neuroimaging data [44]. We aim to test our algorithms' ability to generalize across different studies and populations by employing a dataset aggregated from multiple international sites with different imaging protocols. This approach helps to identify and mitigate any biases that may arise due to specific scanner types, imaging conditions, and demographic characteristics, thereby improving the robustness and applicability of the brain age prediction models in varied clinical and research settings.

It must be noted that the OpenBHB project organizes the validation set (our test set) in a particular way. It is divided into two partitions: the internal test set and the external test set. The former consists of observations from the same sites as the training set, while the latter includes examples from patients treated at five sites not considered in the training set. Both partitions were created using stratified sampling based on age and sex. For the internal set, stratification also considered sites. This particular division of the validation set allows for evaluating the models in scenarios involving both previously seen sites and new, unknown settings, thereby assessing the model's generalization capability on unseen examples.

We utilized $N = 3984$ T1-weighted MRI scans of a cohort of Healthy Controls (HC) in the OpenBHB dataset, collected from 62 international sites (Table 1).

The subjects included in the study came from distinct backgrounds, including European-American, European, and Asian origins, ensuring a varied sample. The overall age and sex distributions for OpenBHB are shown in Fig. 1. The T1-weighted MRI scans were collected using 1.5 Tesla and 3 Tesla scanners, with varying characteristics such as manufacturers and acquisition parameters (e.g., repetition time, echo time, sequence name, flip angle, and acquisition coil). Three modalities derived from the same T1-weighted MRI scans are available: Voxel-Based Morphometry (VBM), Surface-Based Morphometry (SBM), and quasi-raw (minimally preprocessed) data. This study focuses on the quasi-raw T1-weighted images and Surface-Based Morphometry indices derived using FreeSurfer.

3.2 Preprocessing of quasi-raw MRI images

Minimally preprocessed data were generated using a series of neuroimaging preprocessing tools. Initially, ANTS[3] was used for bias field correction to address intensity inhomogeneities in the MRI scans. Next, FSL

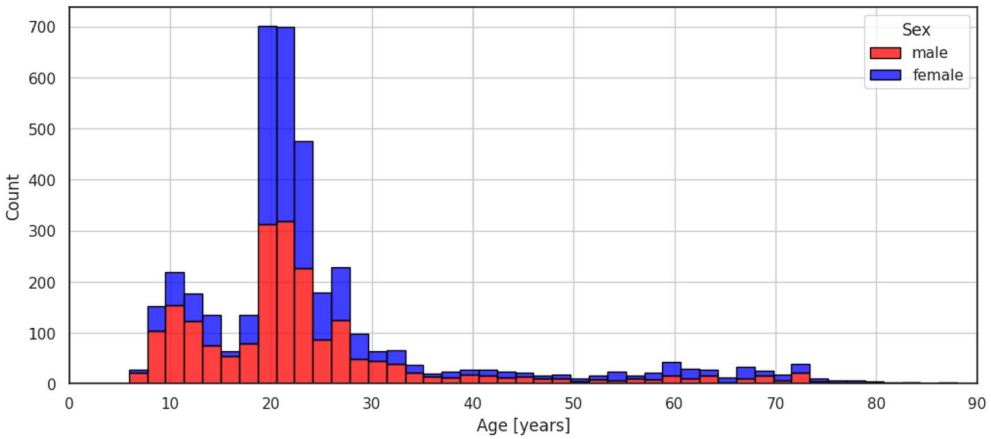


Fig. 1 Overall age and sex distributions of the subjects in the OpenBHB dataset

[3] https://github.com/ANTsX/ANTs.

FLIRT[4] was employed with 9 degrees of freedom (excluding shearing) followed by affine registration to the MNI template. Finally, a brain mask was applied to remove non-brain tissues, ensuring that only brain structures were retained in the final images.

Following these initial steps, the data underwent further preprocessing. First, undersampling was performed, reducing each dimension of the MRI scans by half. This process decreased the resolution, resulting in final dimensions of $91 \times 109 \times 91$ voxels. Undersampling helps in reducing the computational load and storage requirements while preserving essential structural information.

Subsequently, z-score normalization was applied to standardize the intensity values across the dataset. This step involved calculating the mean and standard deviation of the voxel intensities and transforming each voxel value to its corresponding z-score.

3.3 Extraction of morphological features

Morphological features were obtained from the T1-weighted raw scans using the recon-all pipeline from the FreeSurfer software. This pipeline processes and analyzes structural MRI data through a series of steps, including intensity normalization, skull stripping to remove non-brain tissues and segmentation of gray matter and white matter. It also performs hemispheric-based tessellations to create a mesh representation of the cortical surface, followed by topology correction and inflation to visualize cortical folding patterns. The final step registers the cortical surfaces to the "fsaverage" template, a standardized brain template used to facilitate comparisons across subjects.

Based on the data, FreeSurfer derives various morphological measures. In this study, we focused specifically on seven ROI-based features computed using the Desikan-Killiany atlas [45]. These features include cortical thickness (mean and standard deviation), gray matter volume, surface area, integrated mean curvature, integrated Gaussian curvature, and the intrinsic curvature index.

Normalization of volumetric data is a crucial step in neuroimaging studies to mitigate the confounding effects of individual differences in brain size [46]. Intracranial volume (ICV) can vary significantly between individuals due to factors such as age, sex, and overall body size. By normalizing brain volumes, we ensure that comparisons across subjects focus on specific regions of interest (ROIs) rather than being influenced by variations in overall brain size. Notably, cortical thickness measures were excluded from this normalization process, as they are less susceptible to variations in total brain size and do not require such adjustment.

To achieve this normalization, we first calculated the mean ICV across the entire sample to establish a reference point. Subsequently, a regression analysis was performed with ICV as the independent variable and the volume of each ROI as the dependent variable, yielding the B-weight, which quantifies the relationship between ICV and ROI volume. This B-weight was then used to normalize the volume of each ROI for each subject. The normalization was performed using the following formula:

$$\text{Normalized Volume} = \text{Raw Volume} - (B\text{-weight} \times (\text{ss ICV} - \text{mean ICV})) \quad (1)$$

where "ss ICV" represents the single subject's intracranial volume. This normalization adjusted the raw volumes based on each subject's ICV relative to the sample mean ICV.

Finally, all features underwent min-max normalization, which was applied to the seven characteristics across each of the 68 anatomical regions (34 per hemisphere) defined by the Desikan-Killiany atlas [45]. This process rescaled the features to a fixed range of [0, 1], ensuring that all features were on a comparable scale. This normalization step is crucial as it facilitates subsequent analysis and machine learning applications by eliminating the effects of differing measurement scales across features [47].

4 Methods

4.1 Overview of the ML framework

To address our research questions, we implemented a ML framework organized into two distinct pipelines. These pipelines differ primarily in their tuning processes and the types of architectures employed. The first pipeline, referred to as "Pipeline 1", utilizes established CNN architectures widely used in computer vision and brain age prediction tasks. The second pipeline, "Pipeline 2", is centered around training and testing DNNs.

Both pipelines follow the workflow depicted in Fig. 2. The initial phase in each pipeline involves a tuning process, where various architecture configurations are explored using only the training set data to optimize model performance. Although the tuning process varies between the two pipelines, it consistently results in the selection of the most effective model. Once the optimal architecture is identified, it is employed in a Leave-One-Site-Out (LOSO) cross-validation scheme and a standard training-evaluation process.

[4] https://web.mit.edu/fsl_v5.0.10/fsl/doc/wiki/FLIRT.html.

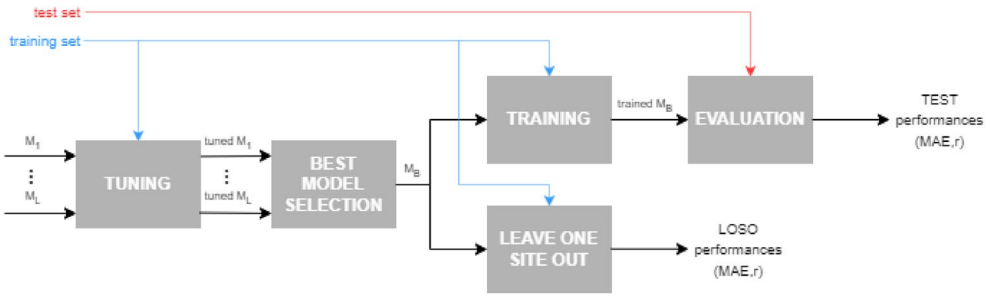


Fig. 2 General workflow of the pipelines

Before delving into the specific differences between the two pipelines, we will first discuss the two steps that are common to both: the *Leave-One-Site-Out* (LOSO) scheme and *Age-Bias Correction*.

The Leave-One-Site-Out scheme (illustrated in Fig. 3) is an iterative process applied exclusively to the training set. Consider a dataset consisting of examples from N different sites. In each iteration of the LOSO scheme, the observations from one specific site are excluded from the training set and designated as the validation fold. A model is then trained on the remaining data from $N-1$ sites and subsequently evaluated on the validation fold, which contains data from the site excluded during training. This approach allows us to rigorously assess the model's ability to generalize to new data, as the validation fold contains examples from a site that the model has never encountered during training. By iterating this process across all sites, we can evaluate the robustness and generalizability of the model across different acquisition settings.

Brain age prediction is typically formulated as a regression problem. A common limitation of regression models is the phenomenon known as regression toward the mean [48, 49], which in the context of brain age prediction is called age-bias. Age-bias manifests as a systematic tendency of the model to underestimate the age of younger individuals and overestimate the age of older individuals. To address this issue, an age-bias correction method [50] has been proposed.

Let Y represent the age predictions generated by our model on the training set, and let X denote the corresponding true ages. Assuming that Y is a function of X, the relationship between the predicted and true ages can be modelled using linear regression, as shown in Eq. 2.

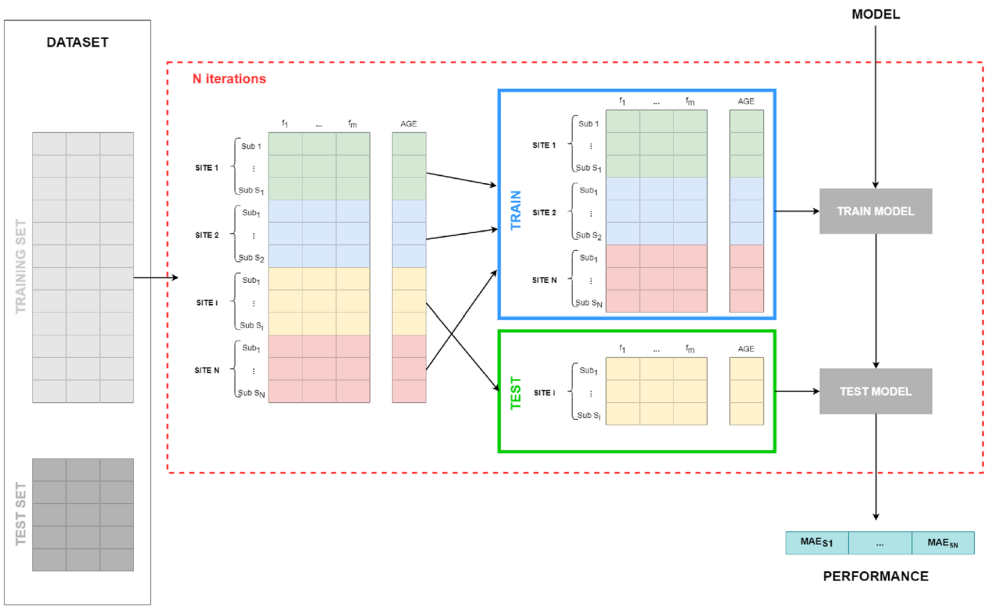


Fig. 3 Leave-One-Site-Out (LOSO) scheme. At each iteration, one site's data is used as the validation set while the remaining data is used for training

$$Y = a \cdot X + b \tag{2}$$

The correction process aims to adjust the predictions and mitigate the age-bias effect, thereby improving the accuracy and reliability of the brain age predictions. The parameters a and b are used to correct the predictions following the formula in Eq. 3, where Y_{corr} is the collection of corrected predictions. Once these parameters are derived from the training set, they can also be applied to correct predictions on an independent test set.

$$Y_{corr} = Y + [X - (a \cdot X + b)] \tag{3}$$

Each pipeline includes a tuning step and a LOSO step. The age-bias correction is applied after the tuning phase, following the training of the optimal architecture configuration on the whole training set. At the conclusion of both pipelines, an explanation module is embedded, which will be discussed in detail later.

The performance of the models was evaluated using the MAE and the correlation coefficient. The MAE is defined as:

$$\text{MAE} = \frac{1}{t}\sum_{i=1}^{t} |\hat{y}_i - y_i| \tag{4}$$

where t is the sample size for the specific test site, y_i is the chronological age, and $\hat{y}_i$ is the predicted brain age.

Additionally, the correlation coefficient between the chronological age and the predicted age of the subjects was computed to assess the performance of the models over the whole dataset. The correlation coefficient is defined as:

$$r = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{\hat{y}})^2}} \tag{5}$$

where $\bar{y}$ and $\bar{\hat{y}}$ denote the sample mean of the chronological age and the predicted brain age, respectively.

4.2 Pipeline 1

In Pipeline 1, we implemented several widely recognized CNN architectures, commonly used in computer vision and brain age prediction tasks. These models were trained using MRI images from the OpenBHB dataset, pre-processed as described in Sect. 3.2. Specifically, this pipeline employed four architectures: an updated version of AlexNet [51], ResNet-18 [52], DenseNet-121 [53], and SFCN [33]. Each architecture was modified and adapted to meet the specific requirements of our study. For all models except SFCN, we developed 3D versions of the architectures as described in the original publications. Additionally, for ResNet, AlexNet, and DenseNet, we introduced a fully convolutional layer with a dropout rate of 0.5 to function as the regression module. Further modifications were made to the SFCN and ResNet-18 architectures to accommodate the resolution of the input images, ensuring optimal performance on the given data. These adjustments were essential to align the models with the unique characteristics of our dataset and the task at hand.

Table 2 presents the complexity of each architecture both before and after the modifications were applied. As shown, the selected architectures provide a range of solutions with varying levels of complexity, enabling a comprehensive investigation across different model complexities.

The first step in the pipeline is the tuning phase. For each architecture, a grid search with k-fold cross-validation was performed to identify the optimal combination of optimizer and learning rate scheduler, both of which are critical choices for ensuring robust model performance. The grid search explored the ADAM and SGD optimizers, as well as two learning rate schedulers: the step scheduler and cosine annealing with warm restarts. The step scheduler was configured with a period of 20 epochs and a decay factor of 0.3. The cosine annealing with warm restarts scheduler was initialized with a 17-epoch period, which doubled at the end of each cycle. Both schedulers were initialized with a learning rate of 0.01, and a minibatch size of 16 was used.

Each point in the grid search was evaluated using a fivefold cross-validation on the training set, with a maximum of 50 epochs per training run and early stopping implemented with a patience parameter of 7 epochs. The model configuration that resulted in the lowest MAE during the k-fold cross-validation was selected as the optimal setup.

Following the tuning phase, the best configuration for each architecture was identified. The four models were then trained on the complete training set for 100 epochs without early stopping, and the model with the lowest MAE on the test set was selected for further analysis.

Finally, the best-performing model was employed in the LOSO cross-validation scheme and subsequently

Table 2 Comparison per architecture of number of parameters between original and adapted version

Architecture	Number of parameters (M)	
	Original	Adapted
AlexNet	60	2.5
ResNet-18	11	33
DenseNet-121	8	11
SFCN	3	1.1

subjected to an explanation analysis to interpret its predictions.

4.3 Pipeline 2

In Pipeline 2, we employed DNN models for brain age estimation based on the extracted morphological features. The input to the DNN comprised 476 features (7 features for each of the 68 ROIs). Two distinct feed-forward configurations were developed for this pipeline: a homogeneous network architecture and a pyramidal network architecture.

In the homogeneous configuration, a uniform number of neurons was maintained across all hidden layers. In contrast, the pyramidal configuration was characterized by a progressively decreasing number of neurons across the hidden layers, starting with a wide initial layer that narrowed in subsequent layers. The neuron arrangements in the Pyramidal configuration followed predefined setups such as (512, 256, 128), (1024, 256, 128), (1024, 512, 128), (1024, 512, 256), and (1024, 512, 256, 128).

Table 3 presents the range of hyperparameters considered during the optimization process for both network configurations. As shown, these configurations provide different levels of complexity, allowing for a comprehensive investigation of model performance across various setups.

The first step in this pipeline was the hyperparameter optimization phase. A nested cross-validation strategy was employed, combining LOSO cross-validation with an inner fivefold cross-validation (k = 5). For each LOSO iteration, a fivefold cross-validation was performed on the remaining training data to identify the optimal combination of hyperparameters. The Random Search method was used to explore a wide range of hyperparameters, including the number of neurons per layer, the number of hidden layers, activation function, learning rate, dropout rate, and optimizer. A total of 60 iterations were conducted for the Random Search, ensuring a comprehensive exploration of possible parameter combinations.

Table 3 Hyperparameter distribution for Random Search in homogeneous and pyramidal networks

Hyperparameter	Homogeneous network	Pyramidal network
Neurons per layer	128, 256, 512, 1024	Decreasing per layer
Number of layers	3, 4, 5	3, 4
Activation function	relu, tanh	relu, tanh
Dropout rate	0.0, 0.3, 0.5	0.0, 0.3, 0.5
Learning rate	1e−2, 1e−3, 1e−4	1e−2, 1e−3, 1e−4
Optimizer	adam, sgd	adam, sgd

Each hyperparameter configuration was evaluated using the nested cross-validation scheme, with a maximum of 50 epochs per training run and early stopping applied with a patience parameter of 6 epochs. The configuration that resulted in the lowest MAE during the inner fivefold cross-validation was selected as the optimal setup for each LOSO iteration.

Following the hyperparameter optimization phase, the best configuration for each DNN model was identified. The selected models were retrained on the complete training set, excluding the data from the left-out site, for 100 epochs without early stopping. The model with the lowest MAE on the validation set for the left-out site was chosen for further analysis.

Finally, the best-performing model for each configuration was identified by selecting the model most frequently chosen across all LOSO iterations. These optimal models were then trained on the entire training set and evaluated on an independent test set, with MAE as the primary performance metric.

4.4 Statistical comparison of the pipelines

A statistical analysis was conducted to evaluate and compare the performances of the two pipelines on both the LOSO validation and an independent test set. To compare the performance of these models, a Wilcoxon rank sum test ($\alpha = 0.05$) was applied to pairs of MAE distributions resulting from LOSO validation and independent test set.

4.5 XAI

In this work, we adopted different types of explanation algorithms: SHAP [54], DeepSHAP [54] and Grad-CAM [26]. The objective of using these methods is to compare their effectiveness in different settings and to assess their reliability.

4.5.1 SHAP-based explanations

SHapley Additive exPlanations (SHAP) is a local, model-agnostic, post-hoc feature attribution method based on Shapley values from cooperative game theory [55]. This method can be applied to any type of model (agnostic) after the training phase (post-hoc) to assign importance scores to individual features in the context of a specific prediction (local). Typically, SHAP is used to generate explanations for models that handle tabular data, such as deep neural networks. When applied to CNNs, SHAP does not assign scores to features but rather to the pixels or voxels of the input image. To facilitate the application of SHAP to CNNs, DeepSHAP is used. DeepSHAP is a pixel-attribution method that estimates the importance

of each pixel (or voxel) in the input image using an approximation inspired by DeepLIFT [56]. SHAP and DeepSHAP methods require a reference dataset, often referred to as the "mask" or "background," to perform their calculations effectively. This background set consists of a sample of observations from the training dataset, which serves as a baseline for interpreting the contribution of individual features to the model's prediction. The use of such a reference is grounded in the need to evaluate the marginal contribution of each feature to a given prediction in a consistent and unbiased manner.

The Shapley value for a particular feature x_i is computed as the average marginal contribution of that feature across all possible coalitions (subsets) of features. Mathematically, the Shapley value ϕ_i for feature x_i can be expressed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{6}$$

where N is the set of all features. S is a subset of N that does not contain feature x_i. $f(S)$ is the model prediction when only the features in subset S are present. $f(S \cup \{i\})$ is the model prediction when feature x_i is added to subset S.

To estimate $f(S)$ and $f(S \cup \{i\})$, SHAP uses a background dataset of representative samples from the training set. This background set acts as a baseline distribution over which the model's output is evaluated. When computing the SHAP values:

- baseline prediction: the value $f(S)$ is obtained by replacing the features not in S with their corresponding values from the background dataset, effectively marginalizing them out. This provides the model's output when features in S are present, but others are not considered.
- Marginal contribution: to compute the marginal contribution of x_i, the difference $f(S \cup \{i\}) - f(S)$ is evaluated. The feature x_i is included, and its value is compared against the background data to understand its effect on the model's output.
- Averaging over samples: by using multiple samples from the background dataset, SHAP averages the marginal contributions across different representative data points, ensuring the robustness and stability of the explanation.

In practice, a background dataset (mask) is selected that captures the typical distribution of the training data. Each feature's contribution is then evaluated by replacing its value with the corresponding values from the background set, simulating scenarios where the feature might be missing or irrelevant. This method ensures that the importance scores reflect realistic and diverse input conditions rather than overly specific to individual data points.

4.5.2 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) is a widely used post-hoc, local explanation method specifically designed for CNNs. Unlike SHAP, which is model-agnostic, Grad-CAM is tailored to CNNs and leverages gradient information flowing into the final convolutional layer of the network to produce saliency maps [26]. These saliency maps highlight the regions of the input image that are most influential in making the model's prediction, effectively visualizing where the model is "looking" to make its decision.

The core idea behind Grad-CAM is to use the gradients of a target class c (e.g., brain age in this context) with respect to the feature maps A^k of a convolutional layer to generate a coarse localization map. The Grad-CAM algorithm computes the importance weights α_k^c for each feature map k by performing a global average pooling of the gradients:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{7}$$

where y^c is the score for class c (the output of the model for class c); A_{ij}^k denotes the activation at position (i, j) in feature map k; Z is the number of pixels in the feature map (i.e., $Z = \sum_i \sum_j 1$).

Using these weights, the Grad-CAM heatmap $L_{\text{Grad-CAM}}^c$ is computed as a weighted combination of the feature maps followed by a ReLU activation to focus only on the positive influence of the features:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{8}$$

This heatmap is upsampled to the size of the input image to provide a visual representation of the regions that are most relevant for the model's prediction of class c.

Grad-CAM differs significantly from SHAP and DeepSHAP in its approach. It does not require a background dataset (or mask) for the computation of feature importance scores. Instead, Grad-CAM relies solely on the gradient information and the input image, making the process computationally less expensive and faster than SHAP-based methods. However, Grad-CAM's reliance on gradients means that it may be more sensitive to small perturbations in the input image, potentially affecting stability.

4.5.3 Evaluation of feature attribution axioms

When using feature attribution methods, it is desirable that they adhere to the following three axioms [57]:

- *Local Accuracy* (Additivity): for each observation, the sum of the contributions of each feature plus the average prediction of the model must equal the final prediction.
- *Consistency* (Monotonicity): the attribution of a specific feature should increase or remain constant if its contribution increases or remains constant when the considered model changes.
- *Missingness*: if a feature is missing, its contribution should be zero.

Grad-CAM does not inherently satisfy all these axioms, particularly consistency and local accuracy, which SHAP and DeepSHAP fully meet. Grad-CAM is, however, popular in computer vision due to its efficiency and effectiveness in generating intuitive explanations.

4.5.4 Background sets for comparison

While Grad-CAM does not require a background set for generating explanations, SHAP and DeepSHAP do. Considering $M = 57$ sites in the training set and $N = 757$ test examples, in this study, three types of background sets were utilized to evaluate and compare the performance of these methods:

- random background: a collection of 200 observations randomly selected from the training set.
- Stratified background: for a given test observation with a specific age, four training examples closest in age are selected from each site, resulting in a total of 228 samples.
- Site-specific background: for a test observation with a specific age and site, up to 200 training samples from the same site with the nearest age to the test observation are selected. This background type is only used for test observations from sites that are also present in the training set and have at least 30 observations available. If the number of samples exceeds 200, only the first 200 observations are considered.

These different background settings were employed to investigate the stability and consistency of the SHAP, DeepSHAP, and Grad-CAM methods across various scenarios, enabling a robust comparison of their explanation capabilities.

4.5.5 Explanation workflow and evaluation

For each of the 757 examples in the test set, the first two types of background (random and stratified) were always computed. In contrast, the third type (site-specific) was

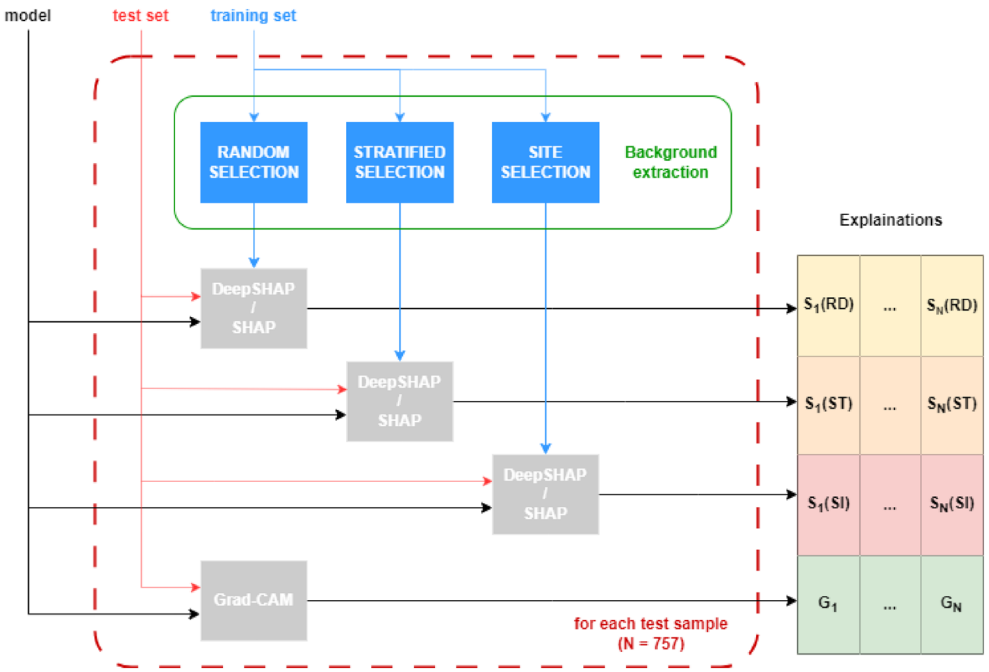


Fig. 4 General workflow of the explanation computation phase: for each test observation, explanations are computed by each method and then collected

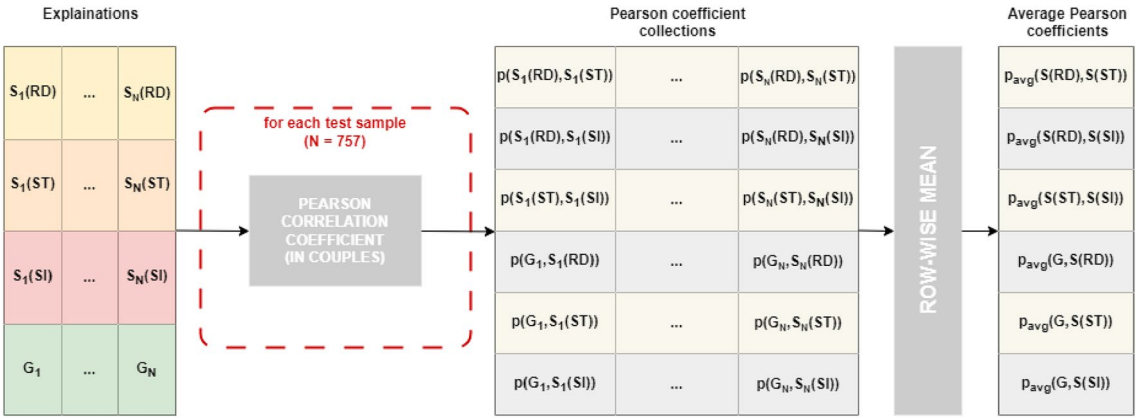


Fig. 5 General workflow of explanation comparison phase: for each observation, correlations in couples of explanation distribution are computed; then, for each couple of methods, the average Pearson is computed

only computed for certain samples from specific sites where sufficient data was available. The general workflow of this explanation phase is illustrated in Fig. 4. Subsequently, the explanations generated by different methods were compared, as outlined in Fig. 5.

In Pipeline 1, the DeepSHAP and Grad-CAM scores were calculated for each test sample using the best-performing CNN model identified during the tuning phase. Both methods produced a 3-D saliency map for each test observation, which was serialized and stored for further analysis. Each test observation resulted in three (or four, when the site-specific background was available) sets of scores. To evaluate the consistency of the explanations provided by different methods, pairwise correlations were computed using the Pearson correlation coefficient ρ, defined in Eq. 5. For each pair of methods, the average correlation ($\bar{\rho}$) and standard deviation (σ_ρ) across all test samples were calculated.

A similar methodology was applied to Pipeline 2, where the optimal model was a DNN. In this workflow, only the SHAP method was used to generate explanations. For each test sample, this resulted in two (or three, depending on the availability of site-specific background) sets of scores. The pairwise correlations of these scores were computed using the Pearson correlation coefficient ρ, and the mean ($\bar{\rho}$) and standard deviation (σ_ρ) were reported for each pair of backgrounds.

This approach allowed us to systematically compare the stability and reliability of the explanation methods across different configurations and backgrounds, providing insights into the effectiveness of each method in generating interpretable and consistent explanations for brain age prediction models.

To assess a proxy of usability of the SHAP outcomes in clinical settings, we investigated the relationship between the SHAP importance scores for each feature and the chronological age of the subjects. This analysis was conducted separately for each pipeline. For Pipeline 1, the correlation analysis was performed using a ROI-based approach. Specifically, the voxel-level importance scores of each subject were averaged within each ROI defined by the atlas. The resulting region-level importance values were then correlated with the chronological age of the subjects. For Pipeline 2, the analysis focused on two distinct subsets of features: one representing the gray matter volume of each region and the other capturing the average thickness of each region. The SHAP importance scores for these features were correlated with the subjects' chronological age in order to evaluate whether the importance assigned to these anatomical attributes by the model corresponds with established age-related changes in the brain. A Bonferroni correction was applied to account for multiple comparisons, with a significance threshold of $\alpha = 0.05$.

To provide a more in-depth exploration of the results, particularly the correlations between SHAP values derived from each model for specific ROIs and the subjects' chronological age, we integrated an interactive dashboard built with Plotly Dash and publicly available at github repository.[5] This dashboard serves as a powerful visual exploration tool designed to help users navigate and interpret complex model outputs. Key features of the dashboard include:

- Manhattan plot: this plot offers a visual summary of the correlation significance between the SHAP values of various cortical thickness ROIs (e.g., "lh-bankssts",

[5] https://github.com/sisinflabaibio/BrainAge-Dashboard.git.

"lh-caudalmiddlefrontal") and the subjects' age. The x-axis represents the features (ROIs), while the y-axis shows the $-log10$ of the p-values, allowing users to quickly identify which ROIs contribute most significantly to brain age prediction across different models. This view aids in comparing how feature importance shifts across different methods and age groups.

- Dependence plot: this plot provides an intuitive way to explore the relationship between the SHAP values (representing feature importance) and the age of the subjects. It allows users to see how variations in specific cortical thickness features affect the model's predictions. Users can dynamically switch between different ROIs to investigate the impact of specific regions on brain age predictions.

By providing these visual representations, the dashboard enables a detailed exploration of how individual brain regions influence model outputs, with a focus on the correlation between SHAP values and subject age for each ROI.

5 Results

5.1 Performance of pipeline 1

The results from the tuning phase of Pipeline 1 on the training set are summarized in Table 4. This table presents the mean and standard deviation of the MAE for each configuration across a fivefold cross-validation for each architecture.

Table 4 Average MAE (and standard deviations) on fivefold cross-validation for each configuration of the CNN architectures in the tuning phase

Architecture	Adam + step[1]	Adam + cos[2]	SGD + step[1]	SGD + cos[2]
SFCN	3.26 ± 0.12	3.18 ± 0.09	4.98 ± 2.20	6.91 ± 3.30
ResNet-18	3.11 ± 0.14	3.31 ± 0.41	3.00 ± 0.18	3.59 ± 1.11
AlexNet	3.63 ± 0.49	3.60 ± 0.74	6.00 ± 3.98	3.52 ± 1.32
DenseNet-121	3.45 ± 0.12	3.16 ± 0.08	2.81 ± 0.10	2.70 ± 0.10

[1] Step scheduler

[2] Cosine annealing with warm restarts scheduler

Notably, ResNet-18 and DenseNet-121 demonstrated stable performance across different configurations, exhibiting higher average performance and lower standard deviations than AlexNet and SFCN. Among the architectures, DenseNet-121 was the most consistent, with the lowest standard deviation, indicating higher precision in its predictions.

The results also suggest that models trained using the Adam optimizer were generally more stable than those trained with the SGD optimizer, as evidenced by lower variance in performance across different folds. This indicates that Adam may be better suited for this particular problem.

For each architecture, the optimal configuration was determined as follows: for SFCN, the best performance was achieved using the Adam optimizer combined with the cosine annealing scheduler. For ResNet-18, the most effective configuration involved using the SGD optimizer with a step scheduler. AlexNet produced the lowest MAE when using SGD with the cosine annealing scheduler; however, although the combination of SGD and the step scheduler resulted in a slightly higher MAE, it exhibited more consistent performance, as indicated by its lower standard deviation. Finally, for DenseNet-121, the best configuration was found to be SGD with the cosine annealing scheduler featuring warm restarts.

Once the best configuration for each architecture was identified, the models were trained on the entire training set and evaluated on the test set. Following this step, the age-bias correction was applied to each model. The results are shown in Table 5. The MAE scores across different models on the overall test set consistently hovered around 3, which is comparable to results reported in the state of the art [33, 58].

It is important to note that the models performed better on the internal partition of the test set, which consists of data from sites included in the training set. In contrast, performance on the external partition-composed of data

Table 5 Performance comparison of different architectures on internal (Int), external (Ext) and global (Glo) test set before and after age-bias correction

	No correction						With correction					
	MAE			r			MAE			r		
	Int	Ext	Glo	Int	Ext	Glo	Int	Ext	Glo	Int	Ext	Glo
SFCN	2.86	4.77	3.86	0.95	0.81	0.87	2.71	4.51	3.64	0.95	0.83	0.88
ResNet-18	2.82	3.93	3.40	0.95	0.89	0.92	2.80	3.74	3.29	0.96	0.90	0.92
DenseNet-121	2.77	3.86	3.34	0.95	0.90	0.92	2.59	3.54	3.08	0.96	0.91	0.93
AlexNet	2.77	4.11	3.47	0.95	0.89	0.91	2.59	3.62	3.13	0.96	0.90	0.92

from unseen sites-resulted in slightly higher MAE values. Nevertheless, even in the external test partition, the MAE never exceeded 5 years.

The age-bias correction was applied to groups of predictions by estimating the regression parameters based on the training set predictions. In Fig. 6, the predictions by DenseNet-121 on the test set are shown in the predicted age vs. chronological age plane (see Figure 14 in Supplementary Material for other architectures). The plot before correction shows how the model does not suffer deeply from the age-bias, and this observation is also true for the other architectures inspected. The trend is confirmed by Table 5, which presents the performance metrics for each model before and after the correction. Overall, the application of the age-bias correction did not lead to significant improvements in performance. This suggests that the models were not substantially affected by the regression toward the mean issue. However, the correction had a more noticeable effect on simpler models, such as SFCN, where a slight performance improvement was observed. Across all models, both before and after the correction, DenseNet-121 consistently showed the lowest error, making it the selected model for the explainability phase.

5.2 Performance of pipeline 2

For the homogeneous network, the configuration that was most frequently selected as the best-performing model included the Adam optimizer, 128 neurons per layer, 3 hidden layers, the tanh activation function, a dropout rate of 0.0, and a learning rate of 0.001. This configuration was identified as the top performer in 6 out of the 57 sites (Table 6).

For the pyramidal network, the configuration most frequently selected as the best-performing model consisted of the Adam optimizer, a neuron arrangement of (1024, 512, 256, 128) across 4 layers, the ReLU activation function, a dropout rate of 0.0, and a learning rate of 0.0001. This configuration was identified as the top performer in 12 out of the 57 sites (Table 7).

The final performance of the optimal models was evaluated on an independent test set. For the homogeneous DNN, the results were as follows: $MAE = 3.05$ (internal),

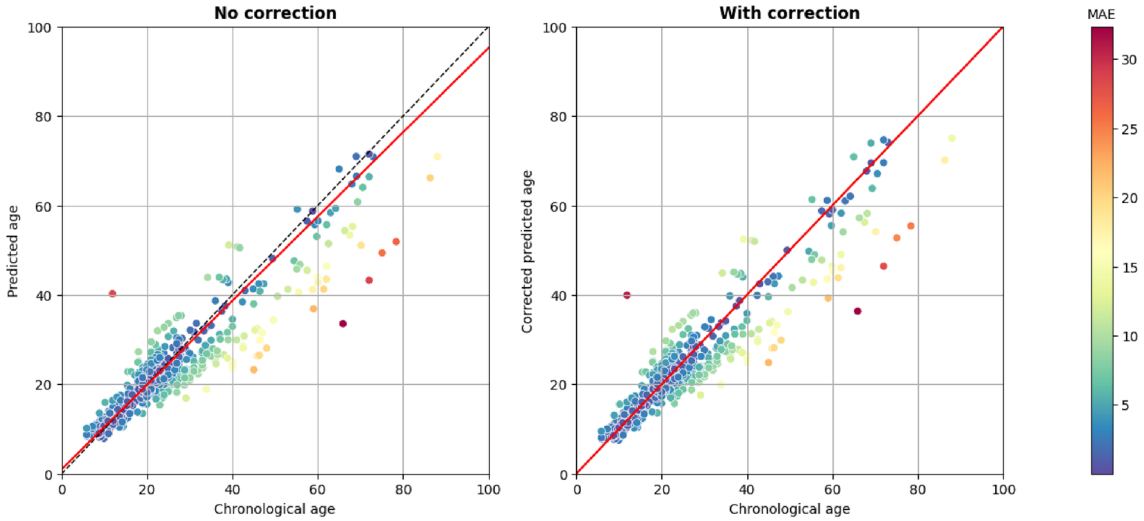


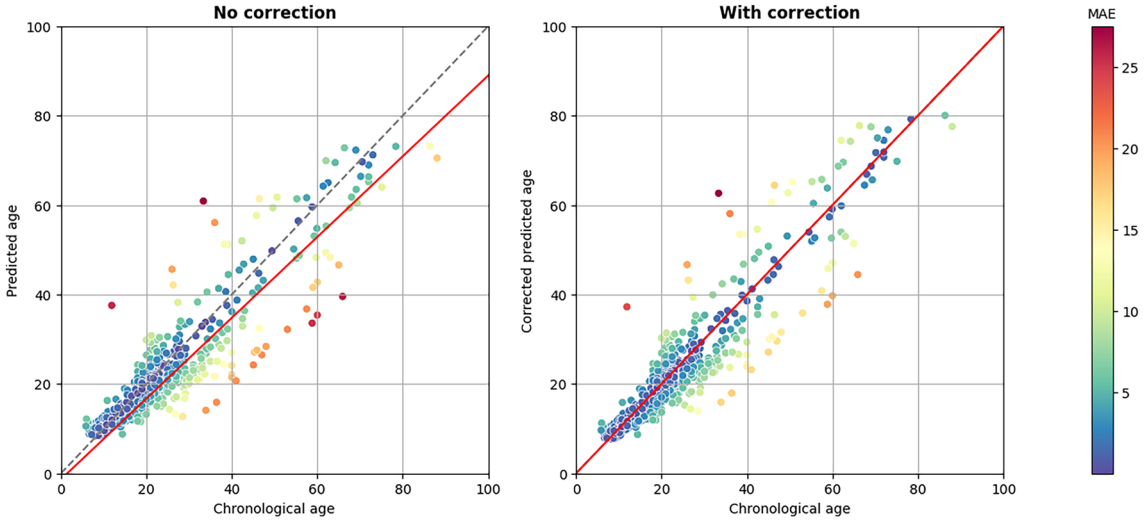
Fig. 6 Predictions on the test set for the *DenseNet-121* model, visualized in the predicted age vs. chronological age plane. In each graph, the data points are scattered around the trend line, with greater deviations observed at older ages. The red line represents the regression line calculated from the training set predictions. The color of the points indicates the MAE, with a gradient ranging from blue (low error) to red (high error)

Table 6 Best configurations for the homogeneous DNN

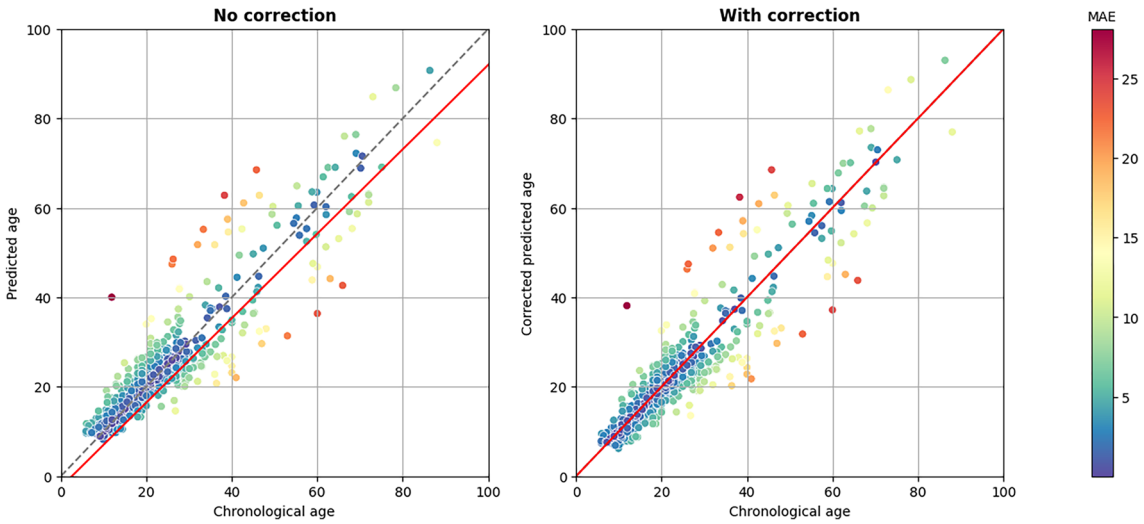
Optimizer	Neurons	Layers	Activation	Dropout rate	LR	Frequency
adam	128	3	tanh	0.0	0.0010	6
adam	1024	5	relu	0.0	0.0001	5
adam	512	5	relu	0.0	0.0001	5
adam	1024	5	tanh	0.0	0.0001	3
adam	1024	4	tanh	0.0	0.0001	3

Table 7 Best configurations for the pyramidal DNN

Optimizer	Neurons	Layers	Activation	Dropout rate	LR	Frequency
adam	1024, 512, 256, 128	4	relu	0.0	0.0001	12
adam	1024, 512, 256	3	tanh	0.0	0.0001	9
adam	512, 256, 128	3	tanh	0.0	0.0001	4
adam	1024, 256, 128	3	relu	0.0	0.0001	3
adam	1024, 256, 128	3	tanh	0.0	0.0001	3



(a) Homogeneous DNN.



(b) Pyramidal DNN.

Fig. 7 Predictions on the test set for the two DNN models in the predicted age vs. chronological age plane. **a** shows the *homogeneous DNN* model, while **b** depicts the *pyramidal DNN* model. In both graphs, data points are scattered around the trend line, with larger deviations observed at older ages. The red line represents the regression line derived from the training set predictions. The color of the points indicates the MAE, with a gradient ranging from blue (low error) to red (high error)

Table 8 Performance comparison with and without correction

	No correction						With correction					
	MAE			r			MAE			r		
	Int	Ext	Glo	Int	Ext	Glo	Int	Ext	Glo	Int	Ext	Glo
Homogeneous	3.05	4.05	3.57	0.94	0.87	0.90	2.72	3.67	3.21	0.95	0.89	0.92
Pyramidal	3.58	3.86	3.72	0.92	0.88	0.90	3.26	3.60	3.44	0.93	0.90	0.91

MAE stands for Mean Absolute Error, and *r* represents Pearson correlation coefficient. Int refers to internal test, Ext refers to external test, and Glo refers to global test

$MAE = 3.57$ (global), and $MAE = 4.05$ (external). In comparison, the pyramidal DNN yielded $MAE = 3.48$ (internal), $MAE = 3.72$ (global), and $MAE = 3.86$ (external). The Pearson correlation coefficients for the Homogeneous DNN were $r = 0.94$ (internal), $r = 0.90$ (global), and $r = 0.87$ (external). For the Pyramidal DNN, the correlations were $r = 0.92$ (internal), $r = 0.90$ (global), and $r = 0.88$ (external).

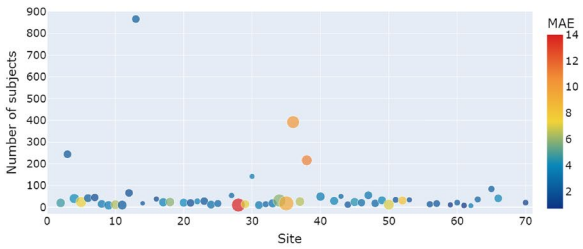
The application of age-bias correction further improved the performance of both models (Fig. 7a and b). For the homogeneous DNN, the corrected results were $MAE = 2.72$ (internal), $MAE = 3.21$ (global), and $MAE = 3.67$ (external). The pyramidal DNN showed improvements to $MAE = 3.36$ (internal), $MAE = 3.44$ (global), and $MAE = 3.60$ (external), as summarized in Table 8.

5.3 Leave-one-site-out

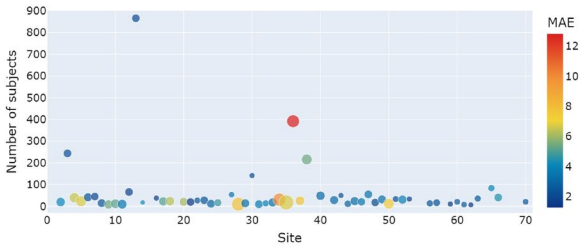
Figure 8a–c present the results of the LOSO evaluation for the DenseNet-121, homogeneous DNN, and pyramidal DNN architectures, respectively. In each bubble plot, the x-axis denotes the site identifier, while the y-axis represents the number of subjects at each site. The size of the bubbles corresponds to the average real age of the subjects at that site. A color gradient from blue to red reflects increasing MAE values, providing a clear visual indication of model performance variations across different sites.

Sites with older subjects tend to exhibit higher MAE values, underscoring the difficulty in accurately predicting brain age in older populations. This challenge is further compounded by the long-tail distribution of subjects aged between 30 and 80 years, resulting in fewer observations within this range. This trend is visible across all three figures, where the bubbles representing sites with older subjects are generally larger and redder, indicating higher prediction errors. Notably, all three models consistently showed higher MAE values for sites 28 (average age: 57 years), 35 (69 years), and 36 (50 years).

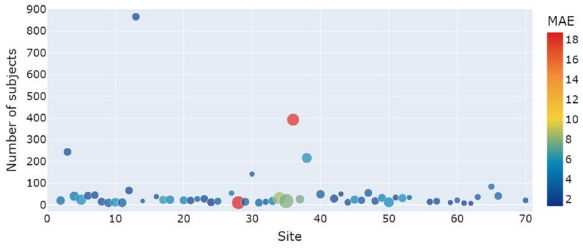
In contrast, sites with younger populations, such as site 5 (average age: 37 years) and site 50 (35 years), were associated with lower MAE values compared to the aforementioned older sites. However, the error remains



(a) DenseNet-121



(b) Homogeneous DNN



(c) Pyramidal DNN

Fig. 8 Bubble plots illustrating the LOSO evaluation for the **a** *DenseNet-121*, **b** *homogeneous DNN*, and **c** *pyramidal DNN* architectures. The x-axis denotes the site identifier, while the y-axis represents the number of subjects per site. The size of each bubble corresponds to the average chronological age of the subjects at that site, and the color gradient, ranging from blue to red, indicates increasing MAE values

Table 9 Mean absolute error per model in leave-one-site-out

Architecture	MAE
DenseNet-121	3.80 ± 2.68
Homogeneous DNN	3.87 ± 2.22
Pyramidal DNN	4.09 ± 3.20

significant, close to 7 years, which is still higher than the average results shown in Table 9. This table reports the average MAE and standard deviation for each architecture across all LOSO iterations. The results indicate that DenseNet-121, the homogeneous DNN, and the pyramidal DNN perform similarly in terms of average MAE, with DenseNet-121 achieving the lowest overall error. Meanwhile, the homogeneous DNN exhibited the smallest standard deviation, indicating more consistent performance. Among the three models, the pyramidal DNN demonstrated the lowest overall performance.

5.4 Statistical comparison of the pipelines

A statistical analysis was conducted on both the LOSO validation and independent test set results to evaluate and compare the performances of the three selected models-DenseNet-121, pyramidal DNN, and homogeneous DNN. The violin plots showing the distributions of the MAE scores for the three architectures for both the LOSO validation and the independent test are shown in Figure 13 in the Supplementary Materials.

The results of the Wilcoxon test on the LOSO validation indicated no statistically significant differences between the distributions of DenseNet-121 and Homogeneous DNN ($p = 0.3691$). This suggests that the performance of the deep learning pipeline and the feature engineering-based approach might not significantly differ in the LOSO validation. Additionally, the pyramidal DNN model showed comparable performance to DenseNet-121 ($p = 0.4769$) and homogeneous DNN ($p = 0.8672$) in the LOSO analysis, further supporting the conclusion that no substantial performance gap exists between these pipelines under this validation scheme.

A similar approach was applied to the error distributions obtained from the independent test set. The Wilcoxon rank sum test revealed that the pyramidal DNN distribution differed significantly from DenseNet-121 ($p < 0.0001$) and homogeneous DNN ($p = 0.0164$), suggesting that the pyramidal DNN model is statistically less accurate on the independent test set. As shown in Tables 5 and 8, the pyramidal DNN consistently exhibited lower performance compared to the other two models. In contrast, no significant difference was observed between the DenseNet-121 and homogeneous DNN models ($p = 0.1196$), indicating that these two pipelines achieve similar performance on the independent test set.

5.5 XAI

Before presenting the results of the explanation analysis, it is important to address a methodological adjustment made for the application of Grad-CAM. This adjustment became necessary after DenseNet-121 was selected as the best-performing CNN model. According to the creators of Grad-CAM [26], the last convolutional layer of a CNN encoder typically offers the best balance between abstract feature representation and spatial detail despite the lower resolution compared to the original input image. Grad-CAM generates a saliency map with the same resolution as the convolutional layer to which it is applied.

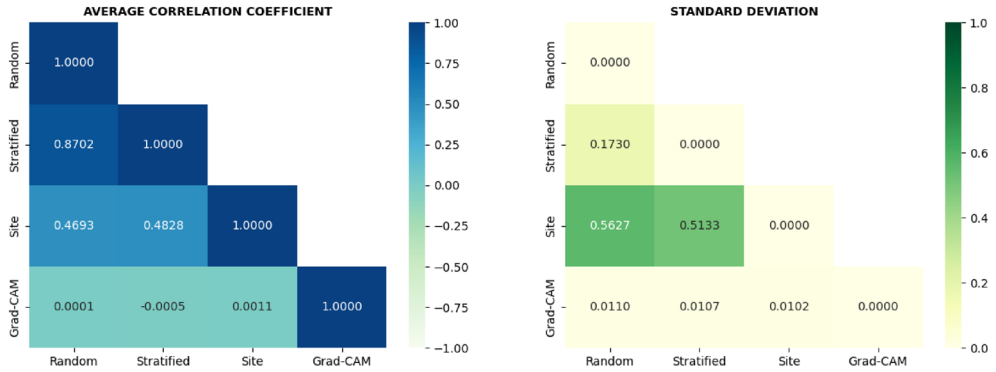
In the case of DenseNet-121, the last convolutional layer had a resolution of $2 \times 3 \times 2$, which is considerably smaller than the original input dimensions of $91 \times 109 \times 91$. Consequently, this layer did not sufficiently preserve the spatial information of the original image, making it unsuitable for meaningful interpretation. Since the last convolutional layer of DenseNet-121 corresponds to the output of the fourth dense block, we opted to apply Grad-CAM to an earlier layer. However, even the third dense block had a resolution of only $5 \times 7 \times 5$, which was still too low for reliable interpretation. To achieve a more appropriate resolution for saliency maps, we examined several layers, including the input convolutional layer ($46 \times 55 \times 46$) and the final layers of the first ($23 \times 28 \times 23$) and second ($11 \times 14 \times 11$) dense blocks. In line with standard practices in Grad-CAM, we averaged the saliency maps computed from these convolutional layers. This averaging approach captures the importance of each voxel across multiple layers, offering a more comprehensive explanation. For each test sample, we generated upsampled maps from the convolutional layers, starting from the input layer and continuing through the final layer of the second dense block, and then averaged them.

The results of the explanation analysis are visualized in Fig. 9a–c, which show the average Pearson correlation coefficients and standard deviations for each pair of methods across different models. Only Pearson correlations with a p-value smaller than $\alpha = 0.05$ were considered statistically significant and included in the analysis.

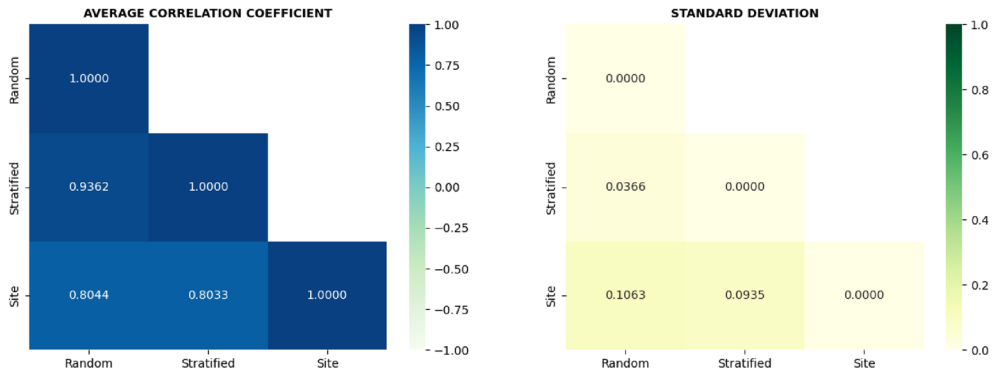
For the DNN models (Fig. 9b and c), the SHAP distributions using different background sets exhibited a strong positive correlation across all comparisons, with correlation values consistently ranging from 0.80 to 0.95. This suggests a high degree of stability and consistency in SHAP's behavior, irrespective of the background used.

In contrast, DeepSHAP did not demonstrate the same level of consistency. As shown in Fig. 9a, while there was a strong positive correlation between the "Random" and "Stratified" background distributions ($\bar{\rho} = 0.87$), the correlations involving the "Site" background were weaker. Specifically, the correlation between "Site" and "Random" was $\bar{\rho} = 0.47$, and between "Site" and "Stratified" it was $\bar{\rho} = 0.48$.

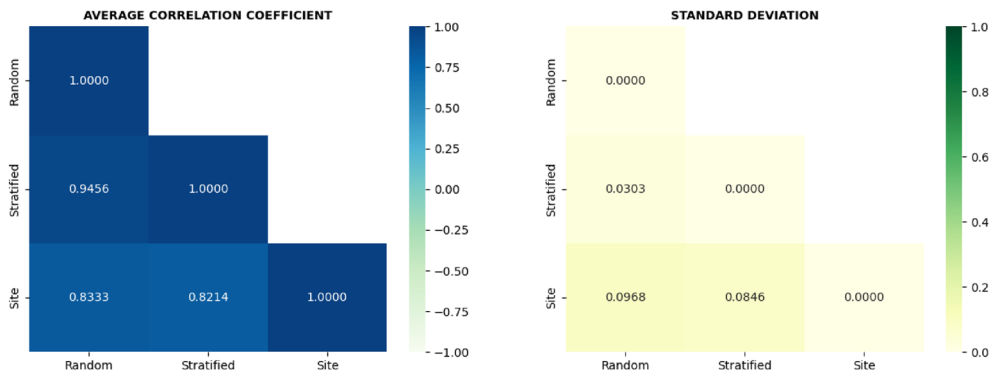
Finally, we evaluated the correlations between Grad-CAM and DeepSHAP with different backgrounds. As depicted in Fig. 9a, there was no meaningful correlation



(a) DenseNet-121



(b) Homogeneous DNN



(c) Pyramidal DNN

Fig. 9 Heatmaps of correlations between explanation distribution for **a** *DenseNet-121*, **b** *homogeneous DNN* and **c** *pyramidal DNN*

between Grad-CAM and DeepSHAP distributions, with $\bar{\rho}$ effectively zero.

The results of the correlation analysis between SHAP importance scores and chronological age are summarized in Figs. 10, 11 and 12. These figures illustrate the relationship between feature importance values and age for each set of features and for each pipeline, providing some insights into the usability of SHAP outcomes in clinical practice.

In particular, Fig. 10 presents the violin plots displaying the distributions of the Pearson correlation coefficients between the XAI importance scores and chronological age for three different explanation types: DeepSHAP (Pipeline 1), SHAP of gray matter volume (Pipeline 2), and SHAP of average thickness (Pipeline 2).

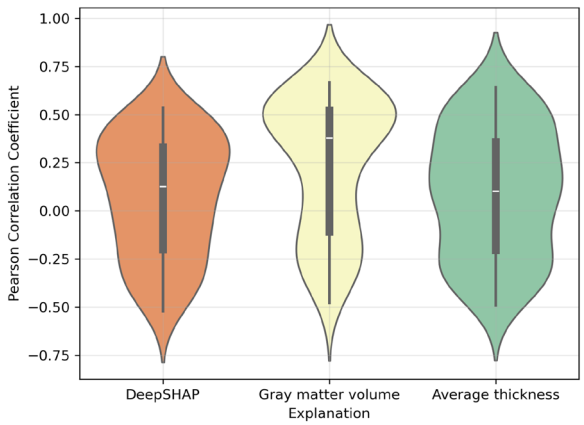
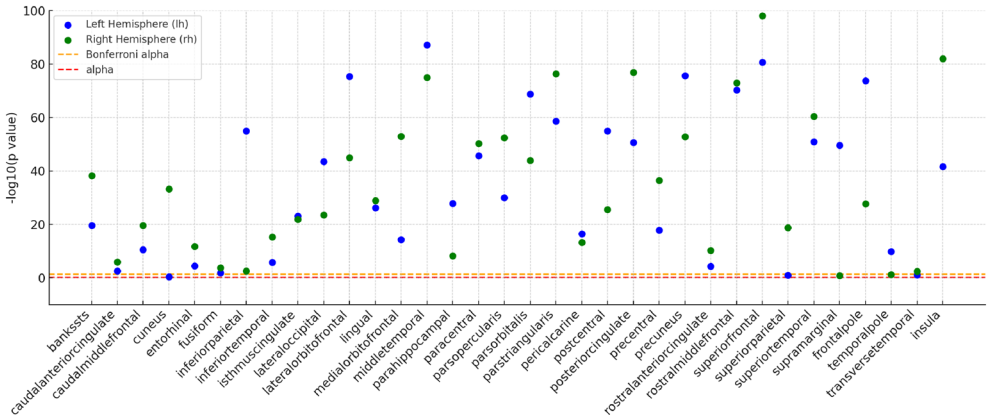


Fig. 10 Violin plots of Pearson correlation score distributions between each feature type and the age of the subjects

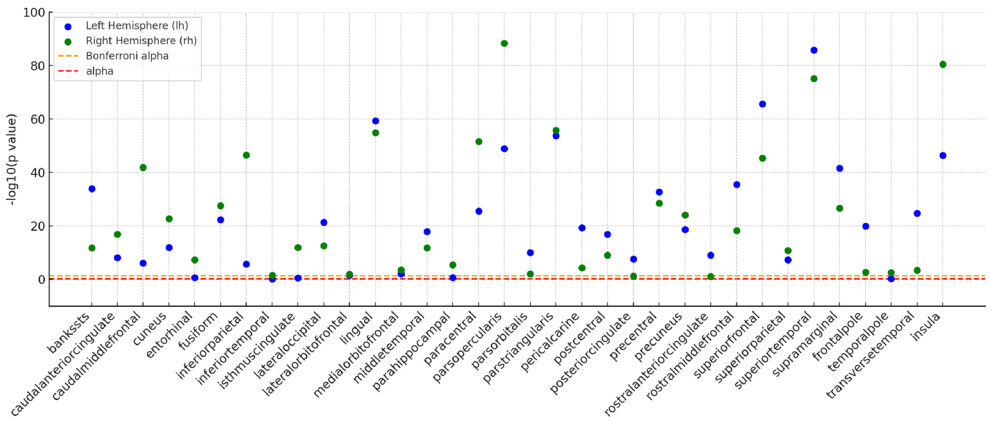
A Kruskal–Wallis test revealed a significant difference in the Pearson correlation coefficients across the groups ($H = 11.20$, $p = 0.0037$), using a significance level of $\alpha = 0.05$.

Post hoc Dunn's tests with Bonferroni correction were conducted to determine which specific groups differed. The results showed that the Pearson correlation coefficients for gray matter volume were significantly different from both average thickness ($p = 0.0277$) and DeepSHAP ($p = 0.0054$). In contrast, no significant difference was observed between average thickness and DeepSHAP ($p = 1.0000$).

The Manhattan plots in Figs. 11a, b and 12 show the correlations between the SHAP values and chronological age across different ROIs for each feature subset. These figures represent static versions of the dynamic plots available in the interactive dashboard. In the dashboard, users can explore these correlations in a more granular and interactive way, allowing for real-time navigation and investigation of specific brain regions.



(a) SHAP (gray matter volume)



(b) SHAP (average thickness)

Fig. 11 Manhattan plots for *SHAP* for **a** *gray matter volume* and **b** *average thickness* features

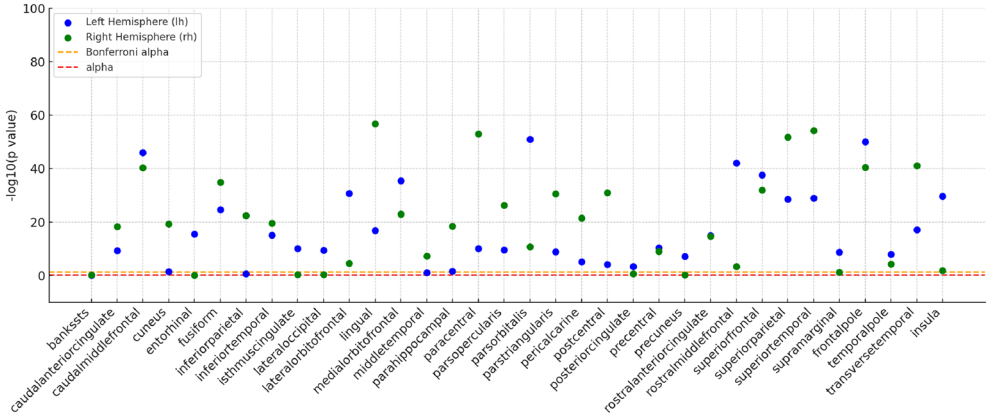


Fig. 12 Manhattan plot for mapped *DeepSHAP* data

- Gray matter volume (Fig. 11a): out of 68 regions, 59 exhibited significant correlations between SHAP importance scores and chronological age after applying the Bonferroni correction ($\alpha = 0.05$). Regions such as the left and the right superior frontal gyrus showed the strongest correlations. In contrast, 9 regions did not show significant correlations, including the left and the right transversetemporal, suggesting that gray matter volume in these areas may be less influenced by age.
- Average thickness (Fig. 11b): in this plot, 54 out of 68 regions were significantly correlated with SHAP importance scores for cortical thickness and chronological age. The most significant correlations were found in the superiortemporal pole. However, 14 regions, including the temporalpole were not significantly correlated, suggesting that cortical thickness in these regions is not strongly linked to aging.
- DeepSHAP (Fig. 12): the DeepSHAP analysis found significant correlations in 53 out of 68 regions. However, 15 regions, such as the left parahippocampal and right precuneus did not show significant correlations, reflecting variability in how DeepSHAP captures the influence of age across different brain regions.

6 Discussion

6.1 RQ1: Do different pipelines yield statistically significant differences in performance?

In this work, we implemented and compared two distinct pipelines for brain age prediction: one utilizing deep learning with 3D CNNs and the other based on feature engineering through FreeSurfer for morphometric feature extraction. To rigorously evaluate the performance of these pipelines on both the LOSO validation scheme and an independent test set, we performed statistical analyses.

For the LOSO validation, the results revealed no statistically significant differences between the performances of the DenseNet-121 and homogeneous DNN pipelines. This suggests that there is no clear advantage in terms of predictive accuracy between using CNNs or morphometric features in this context. This finding is consistent with other studies, such as [59], where the CNN architecture yielded an MAE of 4.006. At the same time, the FreeSurfer-based features resulted in a higher MAE of 5.176, indicating that CNNs can outperform traditional morphometric approaches. However, our statistical tests suggest that for the specific case of LOSO validation, these pipelines might be comparable in terms of overall performance.

On the independent test set, however, our results indicate a statistically significant difference in performance between the pyramidal DNN and the other models, DenseNet-121 and Homogeneous DNN, with the Pyramidal DNN consistently underperforming. Like the LOSO validation, the Wilcoxon rank sum test did not reveal significant performance differences between DenseNet-121 and homogeneous DNN on the test set. This mirrors the conclusions drawn in studies such as [42], where a ResNet model achieved an MAE of 2.85, outperforming a FreeSurfer-based pipeline with an MAE of 4.95. Although the results in both our study and Dufumier's suggest that CNNs are generally more effective for brain age prediction, the statistical analysis in our work shows that CNN and feature engineering pipelines may achieve comparable results under certain conditions.

Our findings are also aligned with those reported in [43], in which MAE values of 4.7 and 3.4 were achieved for CNN and FreeSurfer models, respectively, with

the FreeSurfer-based approach slightly outperforming CNN in that specific context. This exception underscores the importance of dataset characteristics and preprocessing steps in determining the relative effectiveness of each pipeline.

An important consideration that emerged from our analysis is that the choice between a deep learning framework and a feature engineering approach may not solely depend on performance metrics like MAE. Instead, practical factors such as computational requirements and time efficiency play a crucial role. For instance, the FreeSurfer-based feature extraction process can take several hours to a full day for each subject, whereas a trained CNN can process raw images in a matter of seconds. However, training a CNN requires significant computational power, memory storage, and a large collection of training data, making it resource-intensive. Once trained, however, the CNN is highly efficient, whereas the FreeSurfer pipeline must process each new subject individually, leading to slower throughput.

6.2 RQ2: are different XAI methods stable across various parameter settings, and how do they enhance interpretability in the context of these pipelines?

We aimed to evaluate the stability of different XAI methods-specifically SHAP, Grad-CAM, and DeepSHAP-across various parameter settings. Stability is a crucial factor for XAI methods, particularly in clinical applications, where consistency in the generated explanations is essential for building trust with healthcare professionals [60].

Our results showed that the stability of SHAP-based explanations was generally robust across different background datasets, including random, stratified, and site-specific selections. The correlation values for SHAP explanations remained high, typically in the range of 0.80 to 0.95, indicating that the method is highly resilient to changes in background data. This consistency suggests that SHAP can reliably capture feature importance regardless of the background distribution, making it a stable option for explaining deep neural network models in brain age prediction tasks.

On the other hand, the performance of DeepSHAP was more sensitive to variations in background selection. The method displayed weaker correlations when using site-specific backgrounds compared to random or stratified backgrounds, with correlation coefficients sometimes falling below 0.50. This variability suggests that DeepSHAP may not be as reliable as SHAP in certain contexts, particularly when dealing with site-specific data where subtle differences in the dataset characteristics may influence the interpretation.

Grad-CAM, which was applied to the deep learning pipeline, exhibited a different kind of challenge. While the method effectively highlighted important regions in brain MRI images, its outputs were not strongly correlated with SHAP-based explanations. The near-zero correlation values between Grad-CAM and SHAP suggest that these methods provide different information about the models' decision-making processes. This lack of agreement highlights the potential limitations of using Grad-CAM in isolation, as it may not capture the full range of important features that other methods like SHAP can identify.

Overall, the findings indicate that SHAP offers the most stable explanations across different parameter settings and is, therefore, well-suited for applications where consistency is critical. DeepSHAP is a useful tool but may require careful consideration when selecting background datasets, especially in site-specific settings. Grad-CAM, while useful for visualizing important regions in convolutional neural networks, may not be as reliable when compared to SHAP in terms of overall stability and consistency.

6.3 RQ3: how effectively can the combined insights from pipeline performance and XAI explanations support clinical decision-making in brain age prediction tasks?

The final research question aimed to evaluate whether the explanations provided by the selected XAI methods can be effectively integrated into clinical practice. Providing interpretable and trustworthy model outputs is essential when deploying machine learning models in healthcare settings, where decisions directly impact patient outcomes and must be justified to clinicians [61].

Our analysis shows that SHAP explanations, particularly those derived from gray matter volume and average cortical thickness, offer detailed and interpretable insights into the model's decision-making process [38]. SHAP consistently highlighted the most relevant features contributing to brain age predictions, and these explanations were generally aligned with existing clinical knowledge [7, 34, 36, 37, 62]. For instance, SHAP emphasized regions known to be involved in cognitive decline, such as the hippocampus and prefrontal cortex. The feature importance rankings generated by SHAP for individual patients provide actionable insights that can guide personalized treatment plans, making SHAP a strong candidate for clinical adoption. On the other hand, DeepSHAP, while offering similar feature-level explanations, exhibited less consistency across different parameter settings and dataset configurations. Although DeepSHAP was useful in identifying regions such as the entorhinal cortex

and posterior cingulate, which are strongly associated with memory and age-related cognitive changes, the variability observed in some regions raised concerns about its reliability in clinical practice. The inconsistencies between datasets suggest that DeepSHAP may require further refinement or careful parameter tuning before it can be fully trusted in a clinical setting, where reproducibility is paramount.

Grad-CAM, as a visualization tool, provided valuable insights into the regions of the brain that influenced model predictions. However, its coarse, image-level explanations may not always align with the level of detail required in clinical practice [63]. While Grad-CAM effectively highlighted important areas in volumetric brain scans, its limited granularity and the lack of direct feature attribution make it more suitable as a supplementary tool for model interpretability rather than a standalone method for clinical use [64]. Clinicians typically require explanations that offer precise, feature-level insights, and Grad-CAM's region-based heatmaps may fall short of these expectations.

An important consideration when evaluating the clinical utility of these methods is the overlap between regions found non-significant by SHAP (gray matter volume), SHAP (average thickness), and DeepSHAP. Several regions appeared consistently non-significant across all three methods, suggesting that these regions might not be strongly correlated with brain age or that the current models struggle to capture meaningful age-related changes in these areas.

For example, regions such as the left cuneus and right supra-marginal gyrus were not significantly correlated with brain age in both the SHAP gray matter volume and DeepSHAP analyses. Similarly, the right temporal pole was found non-significant in both SHAP (gray matter volume) and SHAP (average thickness) analyses. These overlaps between non-significant regions across methods suggest that these areas may be less sensitive to structural changes related to aging, or that the structural metrics used (e.g., gray matter volume or cortical thickness) may not fully capture relevant age-related variability.

The identification of non-significant regions across multiple methods is informative for clinical practice, as it indicates that these areas may not be reliable markers of brain aging in the current models. Clinicians should exercise caution when interpreting predictions based on regions that consistently show non-significant correlations across different models and methods. This insight reinforces the need for multimodal approaches that combine different brain features with existing structural features to better capture the complexity of brain aging [65].

7 Conclusion

In this study, we developed a systematic framework to compare two distinct pipelines for brain age prediction: a deep learning approach using 3D CNNs and a feature engineering pipeline based on FreeSurfer-derived morphometric features. The framework was applied to a multisite dataset, allowing us to evaluate model performance in a robust context. Our analysis demonstrated that the best models selected from both pipelines performed similarly in cross-validation and on the independent test set.

We also incorporated different XAI methods to assess the interpretability of the models. SHAP emerged as the most reliable tool for providing detailed and clinically relevant explanations, while DeepSHAP showed more variability in its outputs. Grad-CAM, although visually informative, lacked the granularity required for direct clinical use.

One limitation of this work is that we did not conduct a formal comparison of the explanations generated by the different XAI methods. Future studies will address this limitation by formally comparing XAI outputs. Furthermore, we plan to involve clinical users to validate the utility of these explanations in practice, in line with the principles of Human-Centered AI (HCAI). This effort will focus on ensuring that the explanations provided by existing XAI methods are aligned with the needs of clinicians, thus enhancing their utility in making informed decisions. Finally, we will aim to refine the interpretability framework to better integrate multimodal brain features and enhance the clinical relevance of brain age prediction models.

Abbreviations

AD	Alzheimer's disease
MCI	Mild cognitive impairment
MS	Multiple sclerosis
ML	Machine learning
MRI	Magnetic resonance imaging
DL	Deep learning
DNN	Deep neural networks
XAI	EXplainable Artificial Intelligence
SHAP	SHapley Additive exPlanations
Grad-CAM	Gradient-weighted class activation mapping
MAE	Mean absolute error
PLSR	Partial least squares regression
GPR	Gaussian process regression
CNN	Convolutional neural network
ROI	Region of interest
ICV	Intracranial volume
VBM	Voxel-based morphometry
SBM	Surface-based morphometry
LOSO	Leave-one-site-out
BAG	Brain-age gap
AI	Artificial Intelligence
RELU	Rectified linear unit
SGD	Stochastic gradient descent
ADAM	ADAptive moment estimation
HC	Healthy controls
NIfTI	Neuroimaging informatics technology initiative

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40708-024-00244-9.

Supplementary Material 1.

Author contributions

AL conceived the methodological idea for the study. AL, MLNDB and GF defined the methodology and performed the analysis. MLNDB and GF implemented the software. AL, MLNDB and GF wrote the original draft. TDN and EDS supervised the analysis. AL, MLNDB, GF, CA, AF, EDS and TDN analyzed and interpreted the results and edited the final version of the manuscript. All authors have read and approved the final manuscript.

Funding

This work was partially supported by the following projects: "LIFE: the itaLian system wIde Frailty nEtwork", PNRR-MAD-2022-12376656, CTEMT - "Casa delle Tecnologie Emergenti di Matera". "IDENTITA - rete Integrata meDiterranea per l'osservazione ed Elaborazione di percorsi di Nutrizione", and OncologIA.

Data availability

The dataset that supports the findings of this study is publicly available on databases cited in the bibliography.

Materials availability

Not applicable.

Code availability

Pre-trained models will be made available on our GitHub page https://github.com/sisinflabaibio/BrainAge-Dashboard.git following the publication of this paper.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare no competing interests.

Received: 4 October 2024 Accepted: 23 November 2024
Published online: 18 December 2024

References

1. Cole JH, Franke K (2017) Predicting age using neuroimaging: innovative brain ageing biomarkers. Trends Neurosci 40(12):681–690
2. Franke K, Gaser C (2019) Ten years of brainage as a neuroimaging biomarker of brain aging: what insights have we gained? Front Neurol 10:789
3. Cole JH, Franke K, Cherbuin N (2019) Quantification of the biological age of the brain using neuroimaging. Biomarkers of human aging. Springer, pp 293–328
4. Elliott ML, Belsky DW, Knodt AR, Ireland D, Melzer TR, Poulton R, Ramrakha S, Caspi A, Moffitt TE, Hariri AR (2021) Brain-age in midlife is associated with accelerated biological aging and cognitive decline in a longitudinal birth cohort. Mol Psychiatry 26(8):3829–3838
5. Mishra S, Beheshti I, Khanna P (2021) A review of neuroimaging-driven brain age estimation for identification of brain disorders and health conditions. IEEE Rev Biomed Eng 16:371–385
6. Guo X, Ding Y, Xu W, Wang D, Yu H, Lin Y, Chang S, Zhang Q, Zhang Y (2024) Predicting brain age gap with radiomics and AUTOML: a promising approach for age-related brain degeneration biomarkers. J Neuroradiol 51(3):265–273
7. Aghaei A, Ebrahimi Moghaddam M, Initiative ADN (2024) Brain age gap estimation using attention-based resnet method for Alzheimer's disease detection. Brain Inf 11(1):16
8. Franke K, Gaser C (2012) Longitudinal changes in individual brainage in healthy aging, mild cognitive impairment, and Alzheimer's disease. GeroPsych
9. Cole JH, Raffel J, Friede T, Eshaghi A, Brownlee WJ, Chard D, De Stefano N, Enzinger C, Pirpamer L, Filippi M et al (2020) Longitudinal assessment of multiple sclerosis with the brain-age paradigm. Ann Neurol 88(1):93–105
10. More S, Antonopoulos G, Hoffstaedter F, Caspers J, Eickhoff SB, Patil KR, Initiative ADN et al (2023) Brain-age prediction: a systematic comparison of machine learning workflows. NeuroImage 270:119947
11. Bézenac CE, Adan G, Weber B, Keller SS (2021) Association of epilepsy surgery with changes in imaging-defined brain age. Neurology 97(6):554–563
12. Egorova N, Liem F, Hachinski V, Brodtmann A (2019) Predicted brain age after stroke. Front Aging Neurosci 11:348
13. Kaufmann T, Meer D, Doan NT, Schwarz E, Lund MJ, Agartz I, Alnæs D, Barch DM, Baur-Streubel R, Bertolino A et al (2019) Common brain disorders are associated with heritable patterns of apparent aging of the brain. Nat Neurosci 22(10):1617–1623
14. Van Gestel H, Franke K, Petite J, Slaney C, Garnham J, Helmick C, Johnson K, Uher R, Alda M, Hajek T (2019) Brain age in bipolar disorders: effects of lithium treatment. Aust N Z J Psychiatry 53(12):1179–1188
15. Lombardi A, Amoroso N, Diacono D, Monaco A, Tangaro S, Bellotti R (2020) Extensive evaluation of morphological statistical harmonization for brain age prediction. Brain Sci 10(6):364
16. Baecker L, Garcia-Dias R, Vieira S, Scarpazza C, Mechelli A (2021) Machine learning for brain age prediction: Introduction to methods and clinical applications. EBioMedicine 72:103600
17. Baecker L, Dafflon J, Da Costa PF, Garcia-Dias R, Vieira S, Scarpazza C, Calhoun VD, Sato JR, Mechelli A, Pinaya WH (2021) Brain age prediction: a comparison between machine learning models using region-and voxel-based morphometric data. Human Brain Map 42(8):2332–2346
18. Han J, Kim SY, Lee J, Lee WH (2022) Brain age prediction: a comparison between machine learning models using brain morphometric data. Sensors 22(20):8077
19. Tanveer M, Ganaie M, Beheshti I, Goel T, Ahmad N, Lai K-T, Huang K, Zhang Y-D, Del Ser J, Lin C-T (2023) Deep learning for brain age estimation: a systematic review. Inf Fus 96:130–143
20. Velden BH, Kuijf HJ, Gilhuijs KG, Viergever MA (2022) Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Med Image Anal 79:102470
21. Farahani FV, Fiok K, Lahijanian B, Karwowski W, Douglas PK (2022) Explainable AI: a review of applications to neuroimaging data. Front Neurosci 16:906290
22. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P (2022) Transparency of deep neural networks for medical image analysis: a review of interpretability methods. Comput Biol Med 140:105111
23. Lombardi A, Arezzo F, Di Sciascio E, Ardito C, Mongelli M, Di Lillo N, Fascilla FD, Silvestris E, Kardhashi A, Putino C et al (2023) A human-interpretable machine learning pipeline based on ultrasound to support leiomyosarcoma diagnosis. Artif Intell Med 146:102697
24. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4765–4774
25. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I (2020) From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2(1):56–67
26. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE international conference on computer vision (ICCV). pp 618–626
27. Cole JH (2020) Multimodality neuroimaging brain-age in UK biobank: relationship to biomedical, lifestyle, and cognitive factors. Neurobiol Aging 92:34–42
28. Madan CR, Kensinger EA (2018) Predicting age from cortical structure across the lifespan. Eur J Neurosci 47(5):399–416
29. Guan S, Jiang R, Meng C, Biswal B (2024) Brain age prediction across the human lifespan using multimodal MRI data. GeroScience 46(1):1–20

30. Ray B, Duan K, Chen J, Fu Z, Suresh P, Johnson S, Calhoun VD, Liu J (2021) Multimodal brain age prediction with feature selection and comparison. In: 2021 43rd annual international conference of the IEEE engineering in medicine & biology society (EMBC). IEEE. pp 3858–3864.
31. Aycheh HM, Seong J-K, Shin J-H, Na DL, Kang B, Seo SW, Sohn K-A (2018) Biological brain age prediction using cortical thickness data: a large scale cohort study. Front Aging Neurosci 10:252
32. Lombardi A, Monaco A, Donvito G, Amoroso N, Bellotti R, Tangaro S (2021) Brain age prediction with morphological features using deep neural networks: Results from predictive analytic competition 2019. Front Psychiatry 11:619629
33. Peng H, Gong W, Beckmann CF, Vedaldi A, Smith SM (2021) Accurate brain age prediction with lightweight deep neural networks. Med Image Anal 68:101871
34. Dartora C, Marseglia A, Mårtensson G, Rukh G, Dang J, Muehlboeck J-S, Wahlund L-O, Moreno R, Barroso J, Ferreira D et al (2024) A deep learning model for brain age prediction using minimally preprocessed t1w images as input. Front Aging Neurosci 15:1303036
35. Dinsdale NK, Bluemke E, Smith SM, Arya Z, Vidaurre D, Jenkinson M, Namburete AI (2021) Learning patterns of the ageing brain in MRI using deep convolutional networks. NeuroImage 224:117401
36. Levakov G, Rosenthal G, Shelef I, Raviv TR, Avidan G (2020) From a deep learning model back to the brain-identifying regional predictors and their relation to aging. Human Brain Map 41(12):3235–3252
37. Besson P, Parrish T, Katsaggelos AK, Bandt SK (2021) Geometric deep learning on brain shape predicts sex and age. Comput Med Imaging Graph 91:101939
38. Lombardi A, Diacono D, Amoroso N, Monaco A, Tavares JMR, Bellotti R, Tangaro S (2021) Explainable deep learning for personalized age prediction with brain morphology. Front Neurosci 15:674055
39. Mouches P, Wilms M, Rajashekar D, Langner S, Forkert ND (2022) Multimodal biological brain age prediction using magnetic resonance imaging and angiography with the identification of predictive regions. Human Brain Map 43(8):2554–2566
40. Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, Friedrich CM, Nensa F (2023) Explainable ai in medical imaging: an overview for clinical practitioners-beyond saliency-based xai approaches. Eur J Radiol 162:110786
41. Hu G, Zhang Q, Yang Z, Li B (2021) Accurate brain age prediction model for healthy children and adolescents using 3d-cnn and dimensional attention. In: 2021 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE. pp 800–806
42. Dufumier B, Grigis A, Victor J, Ambroise C, Frouin V, Duchesnay E (2022) Openbhb: a large-scale multi-site brain MRI data-set for age prediction and debiasing. NeuroImage 263:119637
43. Jirsaraie RJ, Kaufmann T, Bashyam V, Erus G, Luby JL, Westlye LT, Davatzikos C, Barch DM, Sotiras A (2023) Benchmarking the generalizability of brain age models: challenges posed by scanner variance and prediction bias. Human Brain Map 44(3):1118–1128
44. Saponaro S, Giuliano A, Bellotti R, Lombardi A, Tangaro S, Oliva P, Calderoni S, Retico A (2022) Multi-site harmonization of MRI data uncovers machine-learning discrimination capability in barely separable populations: An example from the abide dataset. NeuroImage Clin 35:103082
45. Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Elsevier 31(3):968-980.
46. Nordenskjöld R, Malmberg F, Larsson E-M, Simmons A, Ahlström H, Johansson L, Kullberg J (2015) Intracranial volume normalization methods: Considerations when investigating gender differences in regional brain volume. Psychiatry Res Neuroimaging 231(3):227–235
47. Gumaei A, Hassan MM, Hassan MR, Alelaiwi A, Fortino G (2019) A hybrid feature extraction method with regularized extreme learning machine for brain tumor classification. IEEE Access 7:36266–36273
48. Everitt BS (2002) The Cambridge dictionary of statistics, 2nd edn. Cambridge University Press, Cambridge
49. Upton G, Cook I (2008) A dictionary of statistics. Oxford paperback reference. Oxford University Press, London
50. Lange A-MG, Anatürk M, Rokicki J, Han LKM, Franke K, Alnaes D, Ebmeier KP, Draganski B, Kaufmann T, Westlye LT, Hahn T, Cole JH (2022) Mind the gap: performance metric evaluation in brain-age prediction. Human Brain Map 43(10):3113–3129
51. Abrol A, Fu Z, Salman M, Silva R, Du Y, Plis S, Calhoun V (2021) Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. Nat Commun 12:353
52. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770–778
53. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 4700–4708
54. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems. NIPS'17. pp 4768–4777. Curran Associates Inc., Red Hook, NY, USA
55. Shapley LS (1953) 17. A value for n-person games. In: Kuhn HW, Tucker AW (eds) Contributions to the Theory of Games (AM-28), vol II. Princeton University Press, Princeton, pp 307–318
56. Shrikumar A, Greenside P, Kundaje A (2017) Learning important features through propagating activation differences. ICML 70:3145–3153
57. Flora M, Potvin CK, McGovern A, Handler S (2022) Comparing explanation methods for traditional machine learning models part 1: an overview of current methods and quantifying their disagreement. ArXiv arXiv:abs/2211.08943
58. Baecker L, Garcia-Dias R, Vieira S, Scarpazza C, Mechelli A (2021) Machine learning for brain age prediction: introduction to methods and clinical applications. EBioMedicine 72(103600):103600
59. Jónsson BA, Bjornsdottir G, Thorgeirsson T, Ellingsen LM, Walters GB, Gudbjartsson D, Stefansson H, Stefansson K, Ulfarsson M (2019) Brain age prediction using deep learning uncovers associated sequence variants. Nat Commun 10(1):5409
60. Yeh C-K, Hsieh C-Y, Suggala A, Inouye DI, Ravikumar PK (2019) On the (in) fidelity and sensitivity of explanations. Advances in neural information processing systems 32
61. Di Martino F, Delmastro F (2023) Explainable ai for clinical and remote health applications: a survey on tabular and time series data. Artif Intell Rev 56(6):5261–5315
62. Leonardsen EH, Peng H, Kaufmann T, Agartz I, Andreassen OA, Celius EG, Espeseth T, Harbo HF, Høgestøl EA, De Lange A-M et al (2022) Deep neural networks learn general and clinically relevant representations of the ageing brain. NeuroImage 256:119210
63. Qiu Z, Rivaz H, Xiao Y (2023) Is visual explanation with grad-cam more reliable for deeper neural networks? a case study with automatic pneumothorax diagnosis. In: International workshop on machine learning in medical imaging. Springer. pp 224–233
64. Suara S, Jha A, Sinha P, Sekh AA (2023) Is grad-cam explainable in medical images? In: International conference on computer vision and image processing. Springer. pp 124–135
65. Lombardi A, Tangaro S, Bellotti R, Bertolino A, Blasi G, Pergola G, Taurisano P, Guaragnella C (2017) A novel synchronization-based approach for functional connectivity analysis. Complexity 2017(1):7190758

Publisher's Note